UNIT - 02

What is ETL- ETL vs ELT - Types of Data Warehouse - DW design & Modeling - Delivery process - OLAP [Online Analytical processing] - Characteristics of OLAP - OLTP [Online Transaction processing] vs OLAP - OLAP Operations - Types of OLAP - ROLAP vs MOLAP vs HOLAP

TOPIC NAME :- What is ETL :-

- Extract / Transform / Load
- Process of combining multiple source called DW
- Organize raw data & Machine learning (ML)

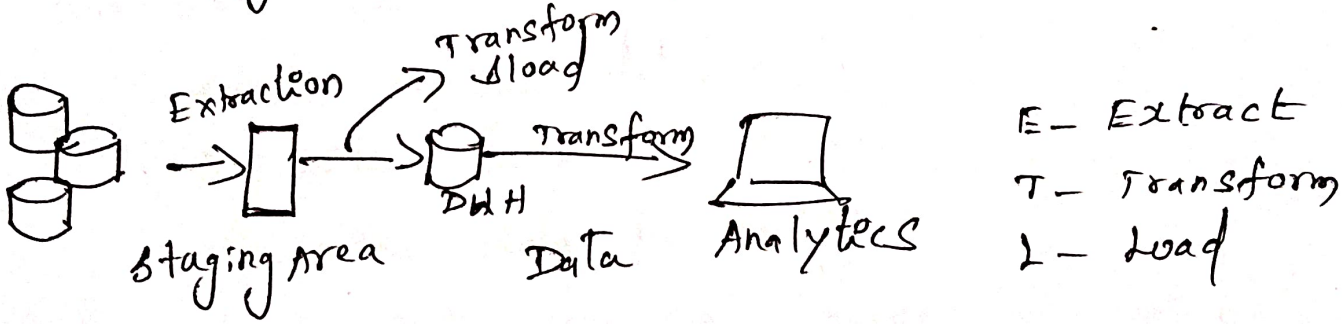


fig :- ETL

Why ETL is important :-

Both structure & unstructure data

CRM- customer Relationship management

IOT - Internet of things

Prepare structure for analytic purpose

Study customer behaviour

ETL Evaluation:-

Convert transaction to relational data formats for analysis

Traditional ETL:-

Support read & write request
Analyze popular items
Identify relationship b/w table/pattern/trends

Modern ETL:-

Create data sinks & receive data from multiple sources

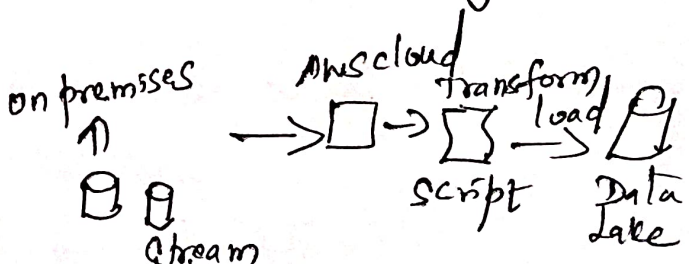
Data Ware houses:-

Store multiple database
Organize table & column
Optimize data processing

Data lakes:- Data like SQL queries/ML/Big data etc

How does ETL work:-

Source DB
Transform data
Load data



dig:- ETL process

EXTRACTION:-

Extract raw data form multiple resources

Happen
- update notification — Extract changes
- Incremental Extraction — notify data record
- Full Extraction — high data transfer

TRANSFORMATION :-

→ consolidate raw data & involve

Basic data transformation - simplify data

Data cleansing — remove error

Data deduplication — remove duplicate record

data format revision — convert into pounds

Advanced data transformation - Easy analysis

Deriving — calculate new value from existing value

Joining - join link & data sources

Splitting — divide target system

Summarization — reduce data volume

Encryption — provide data stream

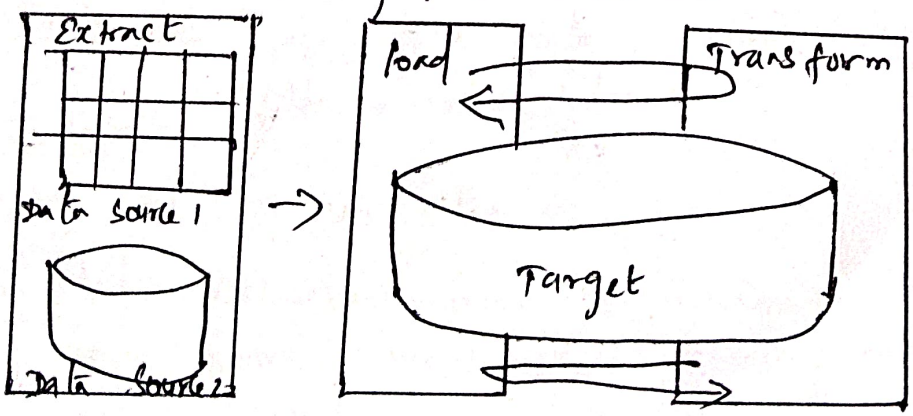
LOADING :- Well defined { full load — load data
Incremental load

Incremental load { Streaming Incremental Load / Batch Incremental load

What is ETL :-

Extract, Load & transform

ETL process - extract data & needs to be loaded repository as well as perform calculations



ELT process

→ Benefits of ELT :-

- Extract, Transform, Load
- Separate transformation step
- Save time & resources
- Easy to change as per requirement

→ ETL VS ELT :-

ELT load raw data to target

ETL – transform secondary processing server

→ Data compatibility :-

represent table rows & columns

handle all type of data like document/images

→ Speed :- difficult to scale / data size increases

→ Cost :- More, simpler data stack & lower setup costs

→ security :- protect data

When to use ETL VS ELT :-

ELT modern analytics

ETL – legacy db, Experimentation, complex analytics

predefined data formats ↓ understand db ↓ improve efficiency & increase performance

IOT Applications :- use sensor data streams

filtered value reduced

Differences :-

A) ETL VS ELT process order

ELT – data extracted from its source

more flexible transformation & easily modified

ELT efficient approach & process target system

B) ETL VS ELT - MAINTENANCE

run efficiently for monitoring / scheduling / Backup / data quality
ensure target system for data transformation

C) ETL VS ELT : COST

data egress cost / less data transforms

D) ETL VS ELT : Security

Perform separate servey
access only specific sources
less secure

E) ETL VS ELT : Hardware requirements

powerful & robust
handle large volume of data
manage data & specialized ETL tool used

F) ETL VS ELT : Support for Data lake

hold large amount of structured data
cost effective, more complex setup
better suited for lake data environment
allow more flexible & scalable

G) ETL VS ELT - Support for Data Warehouse

deal less data cloud lakes & properly structured one

H) ETL VS ELT - performance

Ensure large data volumes & data transformation is parallel
prefer less data size

I) ETL VS ELT : Data movement

Intensive process, reduce amount of data

I) ETL VS ELT : FLEXIBILITY & SCALABILITY

Traditional data processing method

Easy to scale, depend on specific need of requirement

J) ETL VS ELT: SUPPORT FOR UNSTRUCTURED DATA

ETL/ ELT tools used in business process

Incorporate user requirements

K) ETL VS ELT : Data latency

Extract source data, result in highes data latency

for destination lower data latency occur.

Efficient but affect overall data latency of both process

L) ETL VS ELT : DATA LOSS:-

Cause data loss

ETL store all data into target system

Storage requirements may be higher

M) ETL VS ELT : MATURITY:- Extract data for ETL process

CATEGORY	ETL	ELT
→ Stands for	extract/ transform/load	Extract, load & transform
→ process	load predefined format	load target data
→ T & L location	occurs in secondary process	take place target data
→ Data compatability	structured data	structure/unstructured/ semi structured data
→ Speed	slower	faster
→ cost	depend on ETL tool	depend on ELT used
→ Security	custom application protection	manage data protection

TOPIC NAME:- TYPES OF DATA WAREHOUSES

A) HOST BASED DW
B) LAN BASED WORK GROUP
C) SINGLE STAGE DATA WAREHOUSES
D) MULTI STAGE DATA HAREHOUSES
E) STATIONARY DATA WAREHOUSES
F) DISTRIBUTED DATA HAREHOUSES
G) VIRTUAL DATA WAREHOUSES

A) HOST BASED DW:-

Support robust / reliable
handle centrally / Work group environment
Allow automated extraction & cleaning of data
Enable direct access
Transaction processing oriented applications

Host - based (MVS) DW :-

- reside large volume of dB
- high volume data storage

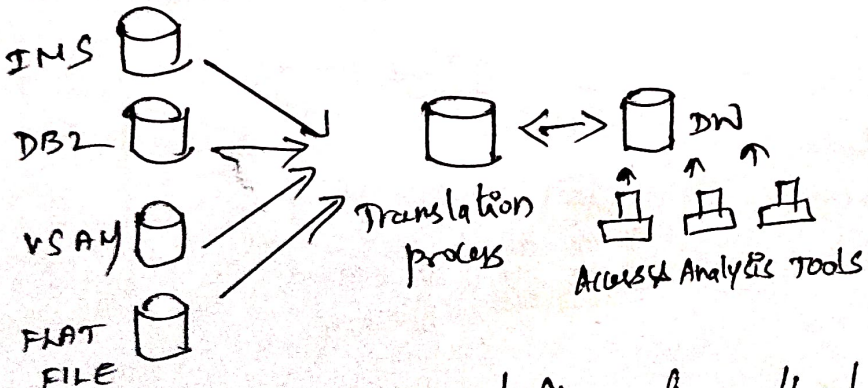


fig: Host based MVS/DW

- follow Unload / transform / load phase
- design, build & maintain / End user access access technique
- support query / report / maintenance for DB2 tool

Host-Based UNIX DATA WAREHOUSES

Extract information UNIX Based dW

Support internetWorking feature

B) LAN BASED WORK GROUP DW:-

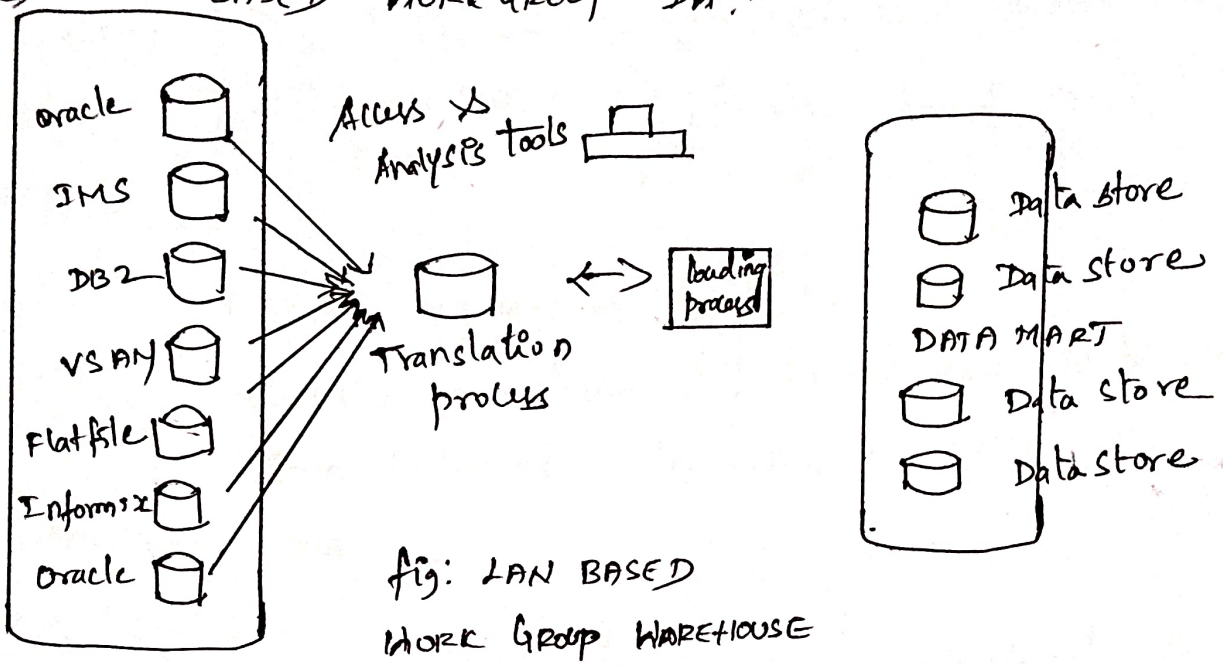


fig: LAN BASED WORK GROUP WAREHOUSE

Extract information multiple LAN

Data Delivery - used in Workgroup level

Maintain LAN environment / often called as data mart.

C) HOST BASED SINGLE STAGE (LAN) DATA WAREHOUSES:-

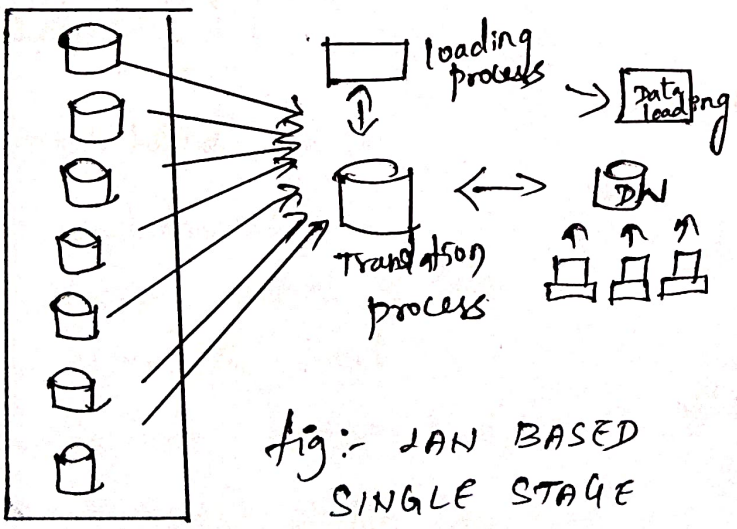


fig:- LAN BASED SINGLE STAGE WAREHOUSE

→ handle Workgroup

→ LAN based solutions

→ dependent one

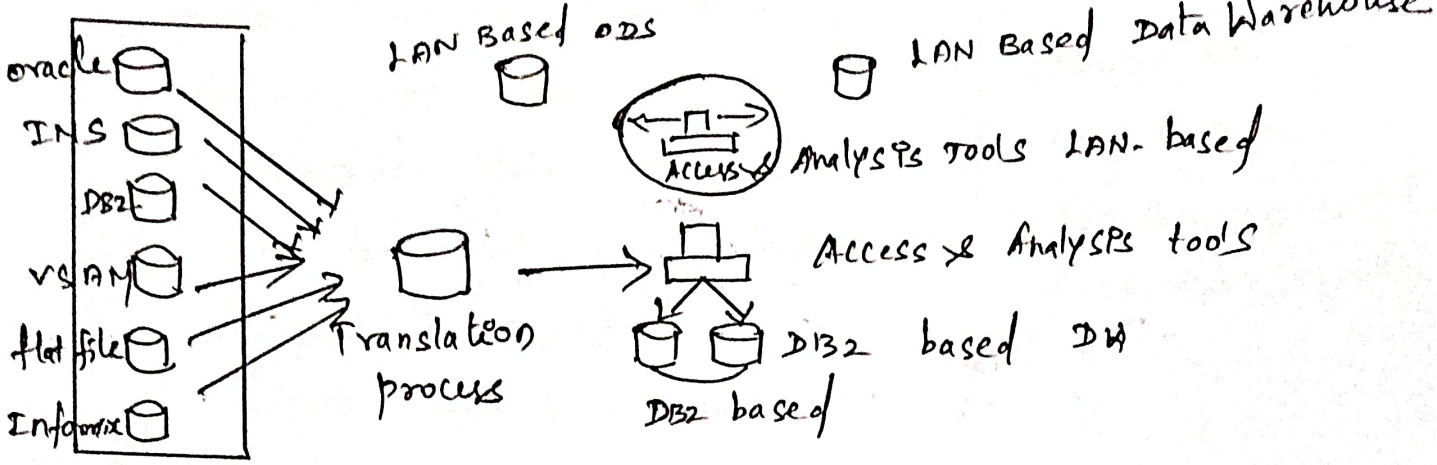
→ design / plan / implementation

→ support DB2 family

→ support business data

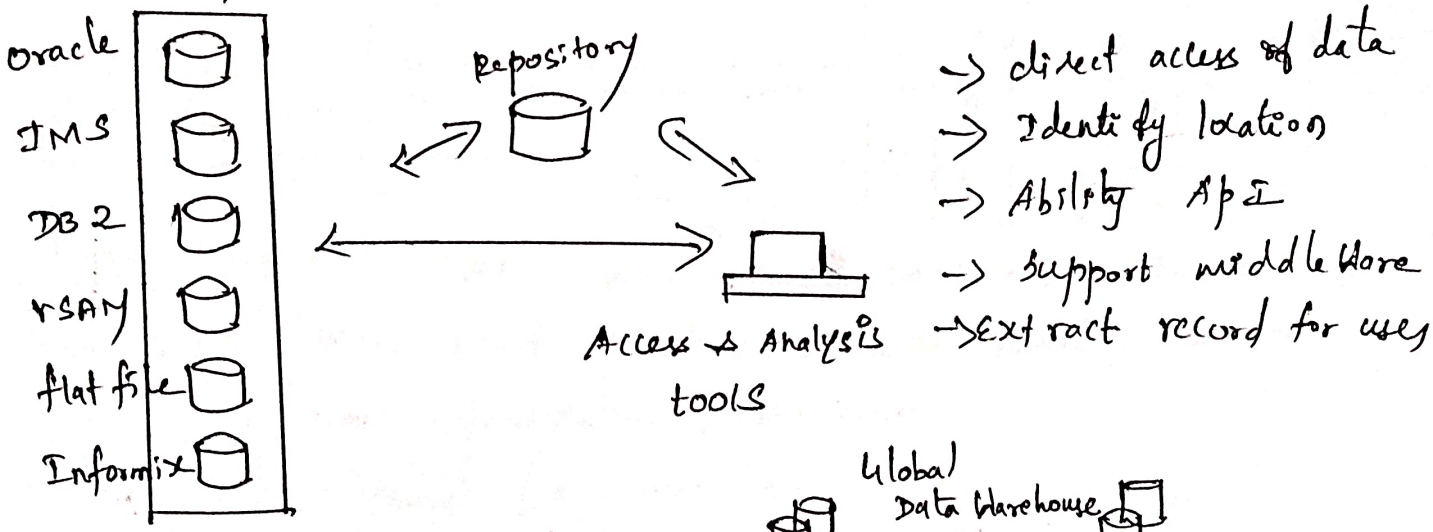
→ feasible

D) MULTI-STAGE DATA WAREHOUSES :-



support aggregation / data mart
suitable end client / historical data to be analyzed
store historical calculation of files

E) STATIONARY DW :-



F) Distributed DW :-

- autonomous
- unique content of data
- record is local
- Include historical data
- Integrate within local site

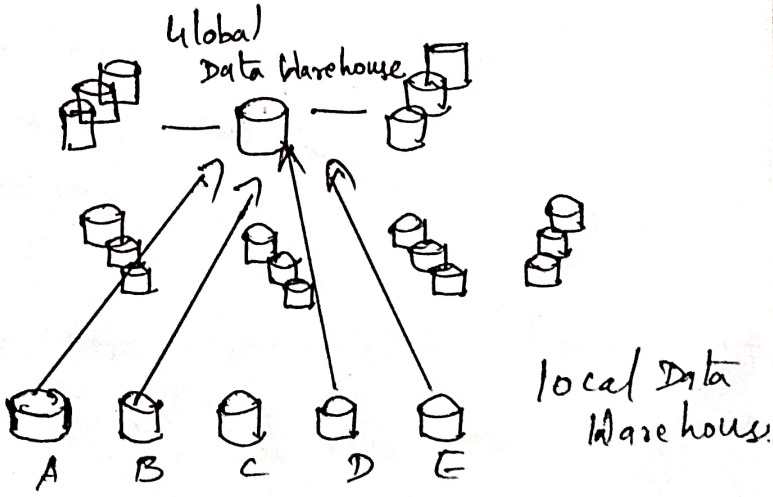
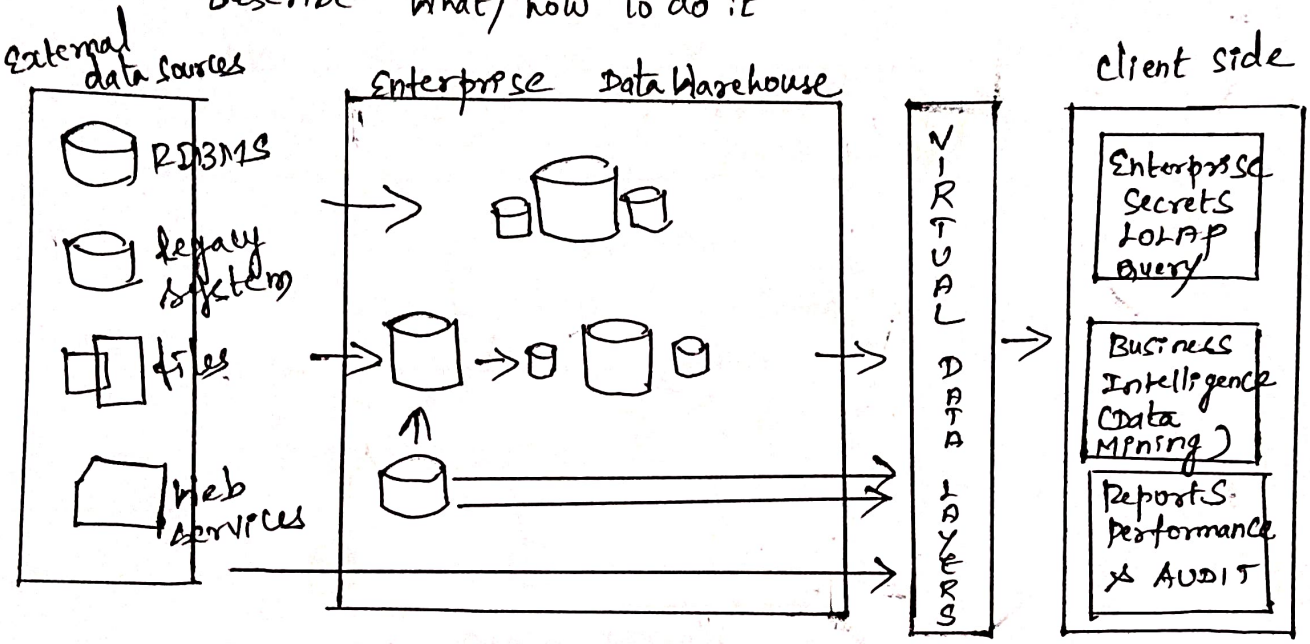


fig: Distributed DW

6) VIRTUAL DH:-

process management facilities
defines end users
Implemented data access network
redundant information loaded
difficult to build
describe What/how to do it

Disadvantage:-
very complex
No summary record
performance degraded



TOPIC NAME: DATA WARE HOUSE DESIGN & MODELING

DESIGN:- OLAP used - online business analytical processing
Meet organization requirement
dynamic/continous/ETL
approach 1) top down
2) bottom up

1) TOP DOWN design approach:-
Subject oriented / time variant/non volatile / data repository
Stores atomic information / data mart apply
Single integrated data source used

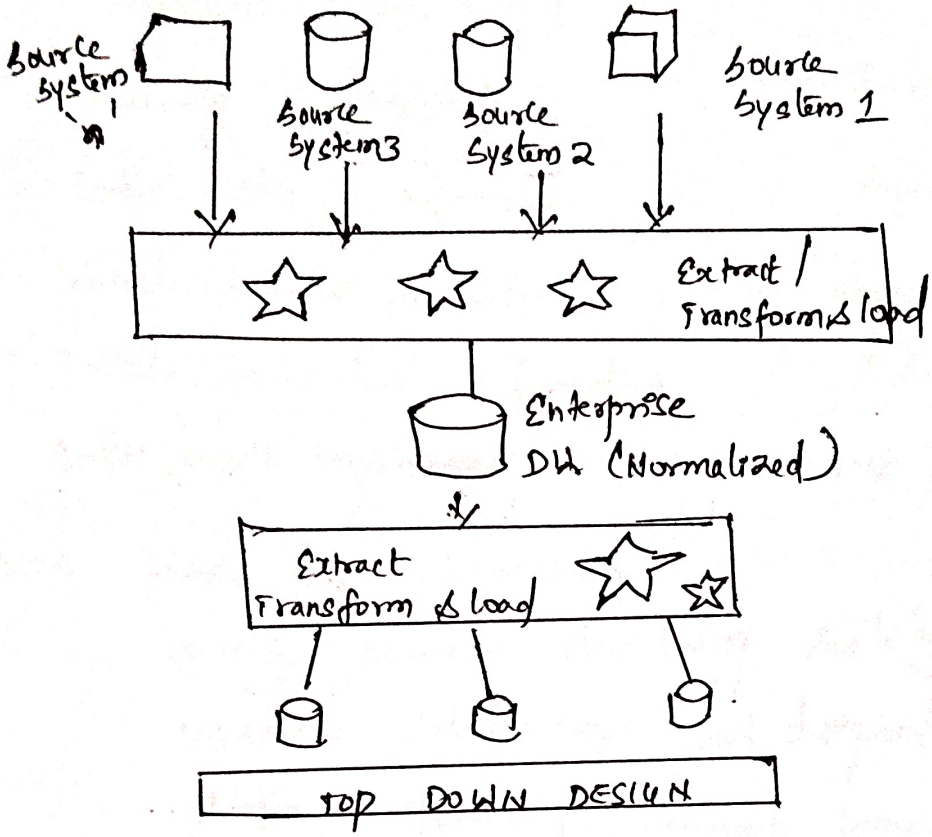
Advantage :-

Easy
Datamart loaded

disadvantage

Inflexible
Cost high

TOP DOWN DESIGN APPROACH

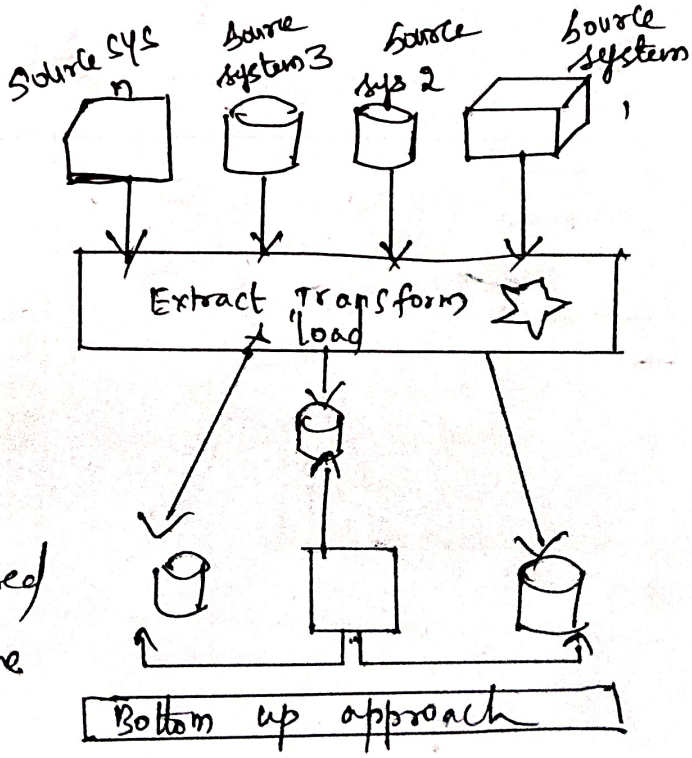


2) BOTTOM UP DESIGN APPROACH :-

query & analyses
business approach
virtual data warehouse
less time / failure less
learn & grow

Adv :- Generate quickly / Datamart used

disadv :- Bottom up approach - less failure

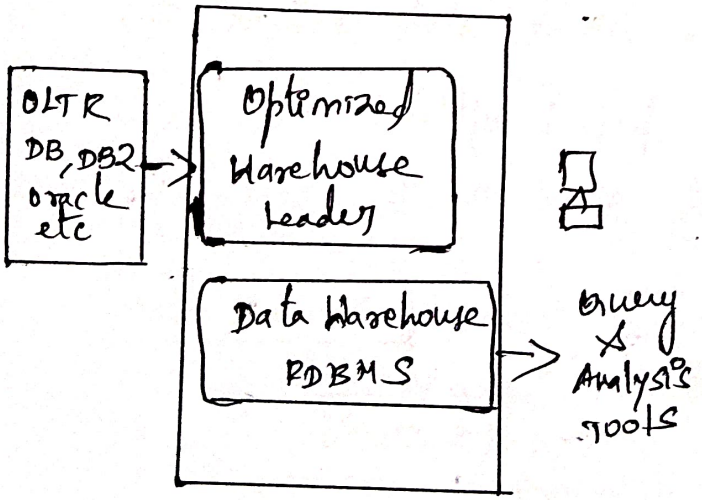


DIFFERENCE :-

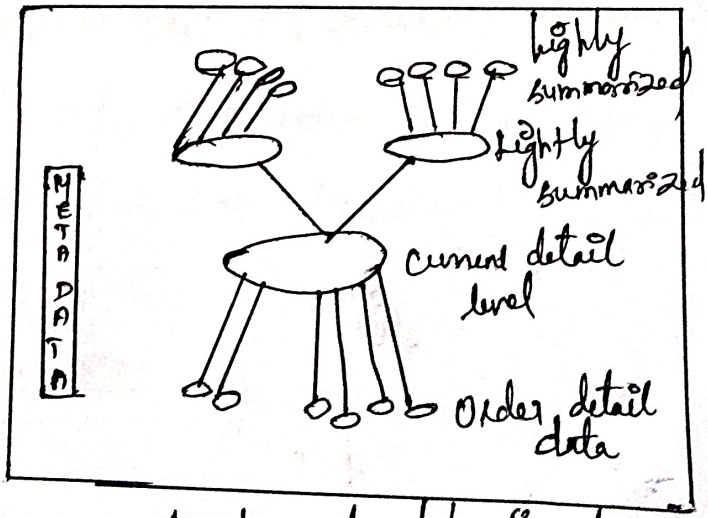
TOP DOWN design	Bottom up design
- Breaks vast problem into smaller subproblems	— Integrate higher one
- Inherently architected	— Inherently Incremental
- Central storage information	— departmental information stored
- Centralized rules & Control	— departmental rules & control
- Include redundant information	— Redundancy can be removed
- quick result implemented	— Less risk of failure

DATA WAREHOUSE MODELING

develop schema, describing reality
visualize relationship, Well designed schema allow
Improve efficiency / support complex queries
Insert / delete / change data



DATA WAREHOUSE MODEL

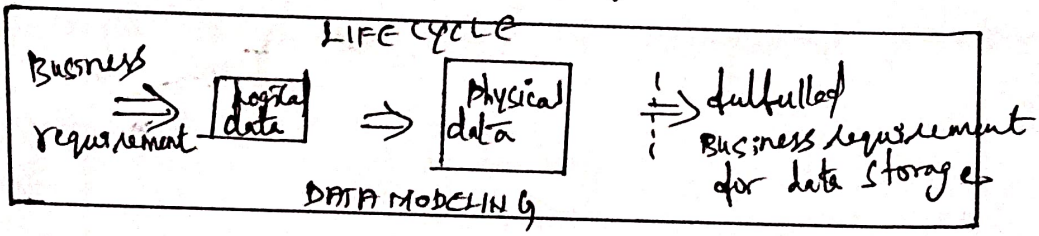


structure of data inside the data Warehouse

dast to access lexpensive / difficult to manage
Older detail data – access current data
Lightly summarized Data – Stored in disk storage / summarize attribute
highly summarized data – outside DW
Metadata – locate data items / map record / summarization b/w current data

DATA MODELING LIFE CYCLE :-

Straight forward proces / fulfill goal for Store, maintain, access

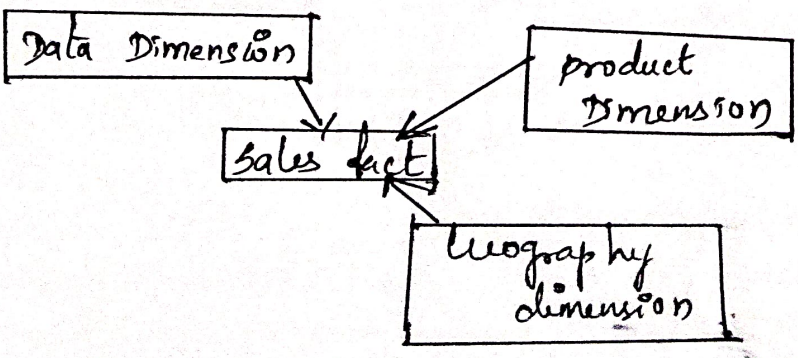


Conceptual Data Model :-

Recognize relationship b/w different entities
Char : entities / relationship
NO attribute specified
NO primary key

Logical data models-

→ Logical data model defines the information as possible
→ NO datatype listed / referential Integrity specified
→ List all attribute for each entity / Normalization.



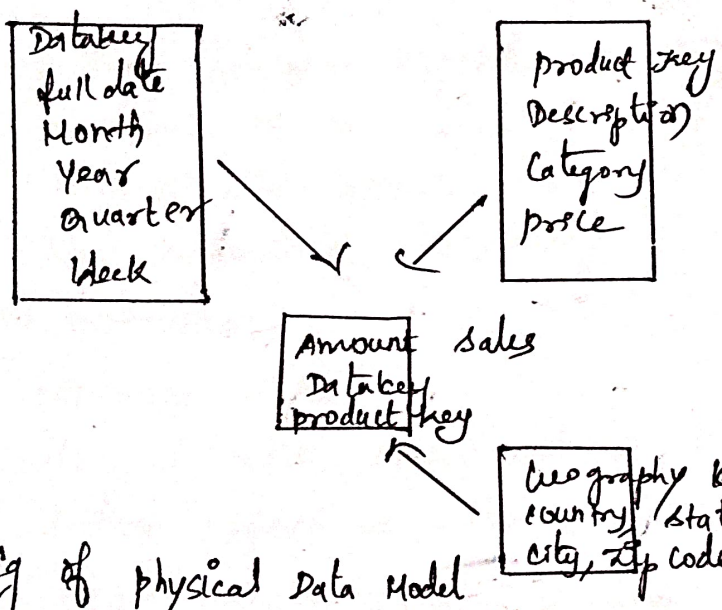
PHYSICAL DATA MODEL :-

- describe how the model is presented
- demonstrate table structures, column names, datatypes
- Store information & include defination of etc new DS
- Enhancing query performance

Characteristics of physical Data Model :-

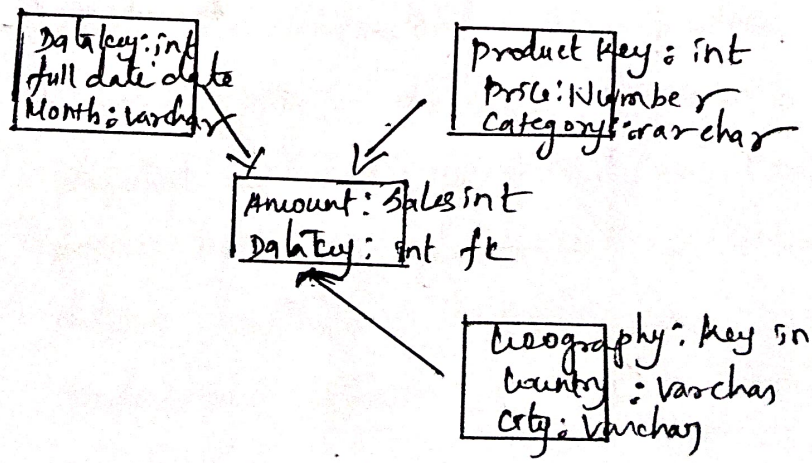
specification all tables & columns

foreign key are used to recognize relationship blw tables



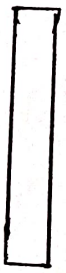
Eg of physical Data Model

TYPES OF DATA WAREHOUSE MODELS :-



Enterprise Warehouse:-
collect record, summarized detailed information
support parallel architecture platforms
used to develop & build
Data Mart:- devided into 2 parts
Independent Data Mart
Dependent Data Mart
VIRTUAL WAREHOUSES:- simple to build / support db servers

TOPIC NAME:- DELIVERY PROCESS:-
Evolve business expands
designed to be flexible
difficult to complete tasks
Not understood completely



Types of DW Models
Enterprise Warehouse
Data Mart
Virtual Warehouse

dig. delivery process

DELIVERY METHOD:-
Minimize risks
does not reduce overall delivery
reduce project & delivery risk

IT strategy:- strategy retain funding for project
Business case: * tends to suffer delivery process
* needs to understand the business case for investment
Education & prototyping:- feasible & promote educational process
defined technical objective / not critical
focus business requirement & technical phase
deliver business benefits
understand short & medium term requirements

BUSINESS REQUIREMENTS:-

provide overall requirement
source system provide data
fulfill short term requirement

Technical Blueprint:-

deliver overall architecture
derive business benefits /data retention /backup & recovery

Building the version:-

Add business benefit

History load:-

create to store / increased data volumes
allow user to analyze recent trends
Identify seasonal trends /complex /perform separate phase

Adhoc query:- generate database query
use access tools

Automation:- fully automated, transforming data into analytics

* Back up, restore, archiving data

Extending scope:-

extend new set of business requirements
using existing information

Requirement Evolution:-

Not static /allow changes reflect within system
opposed to existing queries
Operate pseudo application development process
rework overall system & continually update business needs

TOPIC NAME:-

OLAP - ONLINE ANALYTICAL PROCESSING

OLAP is application architecture

Implement analytical application

using specialized MDDBMS technology

Need for OLAP:-

Array oriented / Multidimensional nature

retrive large number of records

represented relational db & accessed via SQL

Weakness series data & complex functions

OLAP is continuous, iterative & interactive process

OLAP Guidelines:-

does not eliminate methodology

Guidelines :- Multidimensional conceptual view - correspond business pbm

Transparency — proficiency with front end environment tools

Accessibility — perform heterogeneous analysis

Consistent reporting performance — degradation in performance

client/server architecture — Interoperability / flexibility / adaptivity

Generic dimensionality — structure & operational capabilities

Dynamic sparse matrix handling — achieve & maintain performance

Multiuser support — support work group of users

Unrestricted cross dimensional operations — perform rollup calculations

Intuitive data manipulation — drag & drop actions

flexible reporting — analyze analytical process

Unlimited dimensions & aggregation levels — OLAP not impose aggregate levels

TOPIC NAME:- CHARACTERISTICS OF OLAP

A) MULTIDIMENSIONAL DATA ANALYSIS

3D graphics, advance computation & modeling & aggregation functions supports

B) ADVANCED DATABASE SUPPORT :-

Internal & external data sources / drill down & rollups

support very large database

C) EASY TO USE END USER INTERFACES

GUI is simple / useful access

D) SUPPORT CLIENT/SERVER ARCHITECTURE

framework designed / developed / implemented

divide OLAP system / distributed

BASIS	ROLAP	MOLAP	HOLAP
Storage	Relational db used	Multidimension db used	Multidimensional database used
processing	ROLAP slow	MOLAP fast	HOLAP fast
Storage space	large	Medium	small
Storage location	Relational	Multidimensional	Relational
Lateney	low	high	Medium

CATEGORY	OLAP - online Analytical processing	OLTP - online Transaction processing
Definition	db query mgmt syst	db modifying system
Data source	historical data	operational current data
Method	Data Warehouse	DBMS
Application	Data Mining / subject oriented	Business task / application oriented
Normalized	OLAP Not normalized	3NF normalized
Usage of data	plan/pbm solving	day to day operations
Task	multidimension	Business
Purpose	analysis & decision making	Insert/update/delete
Volume of data	Large amount of data	historical data
Queries	slow	fast
update	Not updated	Maintained rigorously
process time	Complex	fast
Types of users	Managed by CEO	Managed by managers
operations	only read / barely write	Both read & write
updates	Batch operation	Brief & quick
Nature	focus Customer	focus market
DB design	focus subject	focus application
productivity	Improve efficiency	Enhance productivity of users

TOPIC NAME:-

20"

OLAP OPERATIONS- ONLINE ANALYTICAL PROCESSING

Execute user queries

SQL based methodology

describe location / time

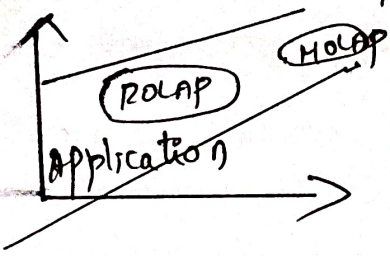
OLAP command → ROLL UP, DRILL DOWN, SLICE, DICE, PIVOT

TYPES OF OLAP

OLAP tool multidimensional db

allow sophisticated user / analyze data

Multidimensional / complex views



A) MOLAP :-

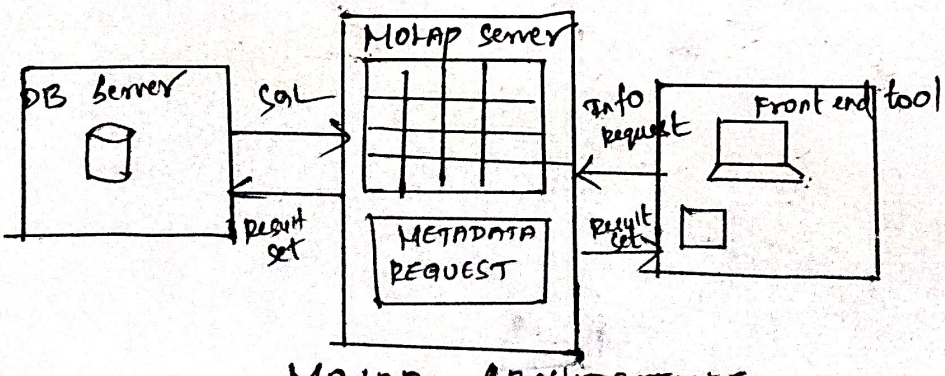
analyze data / aggregated form

predict complex analysis queries

Minimize disk space / perform time series analysis

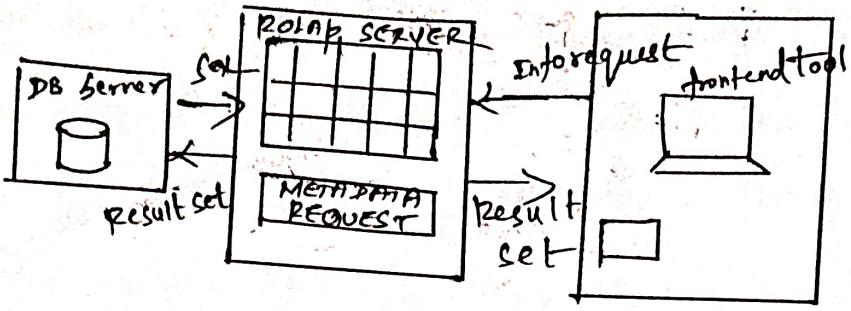
Iterative / comprehensive time series

Access detail data / maintain RDBMS



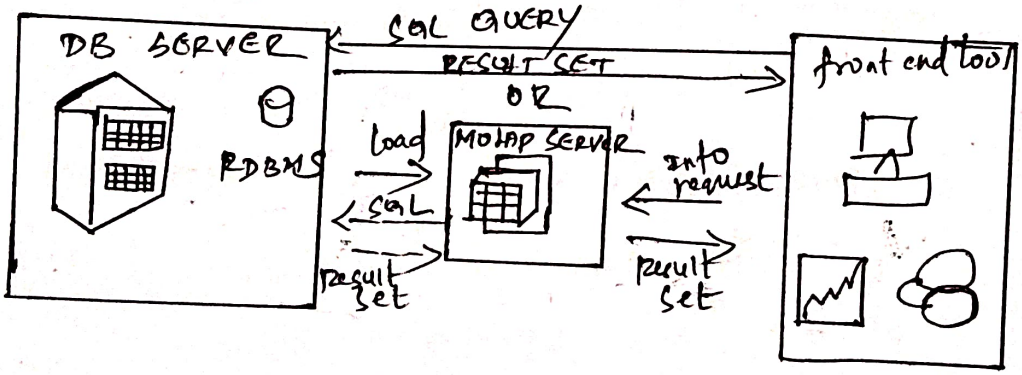
MOLAP ARCHITECTURE

B) ROLAP :- fastest growing style / support RDBMS product
create static multidimensional data structure



ROLAP ARCHITECTURE

C) MQE [MANAGED QUERY ENVIRONMENT]
provide "datacube" & "slice" & "dice" analysis
Store & maintained locally
Support complexity / handle user request
Eg: Sybase / Allow flexibility



Hybrid / MQE - Architecture

— X —